

A mixture model for rare and clustered populations under adaptive cluster sampling

Kelly Cristina M. Gonçalves* and Fernando A. S. Moura†

Abstract

Rare populations, such as endangered species, drug users and individuals infected by rare diseases, tend to cluster in regions. Adaptive cluster designs are generally applied to obtain information from clustered and sparse populations. The aim of this work is to propose a unit-level mixture model for clustered and sparse populations when the data are obtained from an adaptive cluster sample. Our approach considers heterogeneity among units belonging to different clusters. The proposed model is evaluated using simulated data and a real experiment in which adaptive samples were drawn from an enumeration of a waterfowl species in a 5,000 km² area of central Florida.

Keywords: Informative sampling, Poisson mixture, RJMCMC

1 Introduction

In many research studies, it is difficult to observe individuals or collect information from them, such as in surveys of rare diseases, elusive individuals or unevenly distributed individuals. According to [6], rare populations present a few individuals that are sparsely distributed in clusters across a large region. In those cases, the use of conventional sampling methods is not recommended due to the high costs of locating such individuals and the low precision achieved by employing design-based estimators. For instance, suppose that the individuals of interest are spatially distributed in a region upon which

*Departamento de Estatística, Universidade Federal Fluminense (UFF), RJ, Brazil. Email: kelly@est.uff.br

†Departamento de Estatística, Universidade Federal do Rio de Janeiro (UFRJ), RJ, Brazil. Email: fmoura@dme.ufrj.br

we superimpose a regular grid with N cells. Let Y_i denote the grid cell count—for example, the number of endangered plants or animals of interest in the i^{th} grid cell, $i = 1, \dots, N$. The objective is to estimate the population total $T = \sum_{i=1}^N Y_i$. Grid cell sampling methods involve the selection of a subset with $n < N$ grid cells and the observation of the Y_i 's for the selected grid cells. For rare and clustered populations, most of the samples would consist mainly of empty grid cells, yielding poor estimates of T .

To overcome this difficulty, [15] introduced adaptive cluster sampling as a refined method for estimating the size of rare and clustered populations. The scheme is useful for exploring such populations because it allows sampling effort to be focused on the neighborhood of non-empty grid cells in the sample. As stated in [16] adaptive sampling refers to designs in which the procedure for selecting units to include in the sample may depend on values of the variable of interest observed during the survey. For instance, in a survey to assess the abundance of a rare animal species, neighboring sites may be added to the sample whenever the species is encountered during the survey.

Adaptive sampling design starts with an initial probability sample of units, which is selected by a current sample design. Then, when it has found a non-empty grid cell, it also surveys the neighbors of that cell and continues to survey neighbors of non-empty cells until it obtains a set of contiguous non-empty grid cells surrounded by empty grid cells. Selected empty grid cells attract no additional survey effort. This procedure allows the collection of more useful data than simpler sampling methods that ignore the population structure. However, to be effective at a moderate cost, this plan requires some prior knowledge about the structure of the underlying population; see [16] for further details.

For the particular case when the initial sample is a simple random sampling without replacement, [15] derived inclusion probabilities for the networks observed in the sample and used these probabilities to construct design-unbiased estimators of T and their variances. [15] refers to the sets of contiguous non-empty grid cells and their neighboring empty grid cells as clusters. The set of contiguous non-empty grid cells within a cluster is called a network. Empty cells are also defined as networks of size one. The insight in [15] was to base the analysis on networks and to treat the empty edge units of the clusters as unobserved. Adaptive cluster sampling has been performed on real problems

and has been shown to be more efficient than traditional grid cell sampling in different areas. For example, [13] and [8] showed that this method is a viable alternative for sampling forests with rare plants. [14] evaluated the methodology for rare species of waterfowl, and [1] applied it to hydroacoustic surveys in fisheries.

The first attempt to model data obtained by adaptive cluster sampling and to develop a model-based Bayesian analysis was provided by [10]. The use of the Bayesian framework is a natural extension of the key idea behind adaptive cluster sampling, which incorporates the prior knowledge of a clustered population into the inference, as well as into the sampling design. The approach of [10] is based on modeling at the network level. They developed a model for the network counts that considers the informativeness of the adaptive cluster sampling design with respect to the number of counts. However, a crucial aspect of their approach is that, although they do not model the spatial locations of the networks, this decision does not entail any loss of information about the total population because, under the model, the population size does not depend on where the networks are located. They thereby address a potentially difficult problem and are able to proceed relatively simply.

Although the formulation by [10] has certain practical advantages, it does not permit the incorporation of more complex structures, such as spatial dependence between units. Their model supposes homogeneity between all units, even units belonging to different networks, which is equivalent to assuming that the expected total in a network is proportional to its size. However, these assumptions might not be realistic in all real situations.

The aim of this work is to propose a unit-level mixture model for clustered and sparse populations when the data are obtained from an adaptive cluster sample. Our proposed mixture model considers heterogeneity among units belonging to different clusters.

The paper is organized as follows. Section 2 presents the proposed model for estimating the population total of rare and clustered populations from samples selected using adaptive cluster sampling design. It also discusses prior distributions that may be used in this case. The inference specially built for fitting the proposed model is discussed in Section 3, where we also assess the convergence of the MCMC chains by applying informal and formal convergence criteria. Section 4 presents a simulation

study for assessing the estimation of model parameters under different scenarios. It also presents a prior sensitivity analysis of the two possible prior distributions of the parameter that controls the degree of homogeneity among units belonging to different clusters. A comparison of our approach with the one proposed by [10] through design-based and model-based perspectives under different scenarios is presented in Section 5. Finally, Section 6 presents some conclusions and suggestions for further research.

2 A Poisson mixture model for unit counts

The basic mixture model for independent scalar or vector observations Y_i , $i = 1, \dots, n$ is given by:

$$Y_i \sim \sum_{j=1}^k w_j f(\cdot \mid \phi_j), \quad i = 1, \dots, n, \quad (1)$$

where $f(\cdot \mid \phi)$ is a given parametric family of densities indexed by a scalar or a vector ϕ . In general, the objective of the analysis is to make inferences about the unknowns: the number of groups, k ; the parameters ϕ_j 's and the components' weights, w_j , $0 < w_j < 1$, $\sum_{j=1}^k w_j = 1$. The mixture model in (1) is invariant to permutation of the labels $j = 1, \dots, k$. Therefore, it is important to adopt unique labeling to ensure identifiability. For example, we can impose an ordering constraint on ϕ_j 's, such as $\phi_1 < \phi_2 < \dots < \phi_k$.

[17] suggest a Poisson mixture model for dealing with rare events. The interest in this class of models arises here, because it is applicable to heterogeneous populations consisting of groups $j = 1, \dots, k$ of sizes proportional to w_j , from which a random sample may be drawn. The identity of the group from which each observation is drawn is unknown. As stated in [11], due to computational costs, it is natural to regard the group label ϵ_i , for the i -th observation as a latent variable and rewrite (1) as the following hierarchical model:

$$Y_i \mid \phi_j, \epsilon_i = j \sim f(\cdot \mid \phi_j), \quad \text{with } P(\epsilon_i = j) = w_j, \quad i = 1, \dots, n, \quad j = 1, \dots, k.$$

Let us consider a region Ω containing a sparse, clustered population of size T . We superimpose a regular grid on Ω to partition it into N squares. A grid cell is non-empty

if it contains at least one observation and empty otherwise. Let X be the number of non-empty grid cells in Ω . Let $R \leq X$ be the number of non-empty networks, and let $\mathbf{C} = (C_1, \dots, C_R)'$ denote the number of non-empty grid cells within each network, so that $X = \sum_{j=1}^R C_j$. As there are $N - X$ empty grid cells, which are defined to be empty networks of size one, there are $N - X + R$ networks in Ω . Thus, it is possible to extend the R -vector \mathbf{C} to the vector $\mathbf{Z} = (\mathbf{C}', \mathbf{1}'_{N-X})'$ of dimension $N - X + R$, where $\mathbf{1}'_{N-X}$ is the vector of ones with dimension $N - X$. Let $\mathbf{Y} = (Y_1, \dots, Y_X)'$ denote the vector of cell counts, where its elements are the number of observations within each non-empty unit; then, $Y_i \geq 1$. The main goal is to make inferences about the total population $T = \sum_{i=1}^X Y_i$.

The proposed mixture model assumes that the R non-empty network mixture components are heterogeneous, with weights w_j , which in each case are proportional to the number of grid cells inside the networks, C_j . Let us define the latent allocation variable ϵ_i such that $P(\epsilon_i = j) = w_j = C_j/X$, $i = 1, \dots, X$ and $j = 1, \dots, R$.

The mixture model is completed with the hierarchical structure proposed in [10], where they assign distributions to X , R and \mathbf{C} associated with the non-empty grid cells and then, conditionally on the network structure, model the network counts \mathbf{Y} for the non-empty networks.

Our proposed model, can be stated as follows:

$$Y_i \mid \epsilon_i = j, \lambda_j, X \sim \text{independent truncated Poisson}(\lambda_j), Y_i \geq 1, \quad (2a)$$

$$P(\epsilon_i = j) = w_j = C_j/X, i = 1, \dots, X \text{ and } j = 1, \dots, R, \quad (2b)$$

$$\mathbf{C} \mid X, R \sim \mathbf{1}_R + \text{Multinomial} \left(X - R, \frac{1}{R} \mathbf{1}_R \right), \sum_{i=1}^R C_i = X, \quad (2c)$$

$$R \mid X, \beta \sim \text{truncated Binomial} (X, \beta), R = 1, \dots, X, \quad (2d)$$

$$X \mid \alpha \sim \text{truncated Binomial} (N, \alpha), X = 1, \dots, N, \quad (2e)$$

where $\lambda_j / \{1 - \exp(-\lambda_j)\}$ is the mean of the truncated Poisson distribution, and $\mathbf{1}_R$ is the R -vector of ones. It should be noted that, to avoid degeneracy, there is assumed to be at least one non-empty network in the region. Consequently, all the distributions are left-truncated at one.

The distributions stated in (2c), (2d) and (2e) are the same as in the model by [10], but unlike their model, the analysis here is performed at the unit level. In the [10] model, the equations (2a) and (2b) are replaced with independent Poisson distributions truncated at zero: $Y_{.j} \mid \lambda, R, \mathbf{C} \sim \text{Poisson}(\lambda C_j)$, where $Y_{.j} = \sum_{i \in U_j} Y_i$ with U_j denoting the set of units that belong to the network $j, j = 1, \dots, R$. Therefore, our model can handle heterogeneity between units that belong to different networks, which is not considered in the approached proposed by [10].

The selection mechanism that leads to a particular sample $s = \{i_1, \dots, i_m\}$ of size m taken from $N - X + R$ networks is also included in the model and depends only on the network structure, described by X, R and \mathbf{C} . We consider sampling designs whose networks are sampled directly via a sequential procedure where the ordered sample of networks is selected without replacement.

Networks are sampled by the method by probability proportional to size without replacement. Note that the inclusion probability of a network depends on its size Z_i , and the sampling is informative because the components of the random vector \mathbf{Z} are only observed for the sampled networks after being selected. Thus, the probability of selecting the ordered sample $s = \{i_1, \dots, i_m\}$ of m networks must be included in the model likelihood. The joint inclusion probability can be deduced as follows.

Let the event $A_{i_j} = \{\text{the network } i_j \text{ be selected in the } j\text{-th draw}\}$. Thus, the probability of selecting the ordered sample $s = \{i_1, \dots, i_m\}$ of m networks can be written as follows:

$$\begin{aligned} p(s \mid X, R, \mathbf{C}) &= P(\cap_{j=1}^m A_{i_j} \mid X, R, \mathbf{C}) = P(A_{i_1} \mid X, R, \mathbf{C}) \\ &\quad \times \prod_{j=2}^m P(A_{i_j} \mid \cap_{k=1}^{j-1} A_{i_k}, X, R, \mathbf{C}). \end{aligned} \quad (3)$$

Because the networks are sampled without replacement, the conditional probabilities $P(A_{i_1} \mid X, R, \mathbf{C})$ and $P(A_{i_j} \mid \cap_{k=1}^{j-1} A_{i_k}, X, R, \mathbf{C})$ in (3) are, respectively, given by:

$$\begin{aligned} P(A_{i_1} \mid X, R, \mathbf{C}) &= \frac{z_{i_1} \times g_{i_1,1}}{\sum_{i=1}^{N-X+R} z_i - z_{i_0}} \\ P(A_{i_j} \mid \cap_{k=1}^{j-1} A_{i_k}, X, R, \mathbf{C}) &= \frac{z_{i_j} \times g_{i_j,j}}{\sum_{i=1}^{N-X+R} z_i - \sum_{k=0}^{j-1} z_{i_k}}, j = 2, \dots, m, \end{aligned} \quad (4)$$

where $g_{i_j,j}$ is the number of unselected networks of size z_{i_j} after $j - 1$ networks have been selected and $z_{i_0} = 0$.

Substituting the equations in (4) into (3), we finally have:

$$p(s \mid X, R, \mathbf{C}) = \prod_{j=1}^m \frac{z_{i_j} \times g_{i_j, j}}{\sum_{i=1}^{N-X+R} z_i - \sum_{k=0}^{j-1} z_{i_k}}. \quad (5)$$

The sampling procedure entails observing Y_i for the networks in the sample s . The input variables are split into an observed component and an unobserved one, using the subscripts s and \bar{s} , respectively. Thus, we have $X = X_s + X_{\bar{s}}$, $R = R_s + R_{\bar{s}}$, $\epsilon = (\epsilon'_s, \epsilon'_{\bar{s}})'$, $\mathbf{C} = (\mathbf{C}'_s, \mathbf{C}'_{\bar{s}})'$ and $\mathbf{Y} = (\mathbf{Y}'_s, \mathbf{Y}'_{\bar{s}})'$.

As the sampling procedure is informative, it is useful to break the joint probability model into two parts: the model for the underlying complete data, including both observed and unobserved components, and the model for the inclusion probability vector, as stated in (5) (see [7] for further explanation). The complete-data likelihood is defined as the product of these two factors, as stated by [3]. Thus, we can write the complete-data likelihood as

$$\begin{aligned} p(\{i_1, \dots, i_m\}, X, R, \epsilon, \mathbf{C}, \mathbf{Y} \mid \lambda, \alpha, \beta) &= p(\{i_1, \dots, i_m\} \mid X, R, \mathbf{C}) p(\mathbf{Y} \mid \epsilon, \lambda, X) \\ &\quad \times p(\epsilon \mid \mathbf{C}, R, X) p(\mathbf{C} \mid R, X) p(R \mid X, \beta) p(X \mid \alpha) \\ &= \prod_{l=1}^m \frac{z_{i_l} \times g_{i_l, l}}{\sum_{i=1}^{N-X+R} z_i - \sum_{k=0}^{j-1} z_{i_k}} \times \prod_{j=1}^{R_s+R_{\bar{s}}} \prod_{\{i: \epsilon_i=j\}} \frac{\lambda_j^{y_i} \exp(-\lambda_j)}{y_i! [1 - \exp(-\lambda_j)]} \\ &\quad \times \frac{1}{(X_s+X_{\bar{s}})^{X_s+X_{\bar{s}}}} \prod_{j=1}^{R_s+R_{\bar{s}}} C_j^{C_j} \times \prod_{j=1}^{R_s+R_{\bar{s}}} \frac{1}{(C_j-1)!} \left(\frac{1}{R_s+R_{\bar{s}}} \right)^{C_j-1} \\ &\quad \times \frac{1}{(R_s+R_{\bar{s}})!} \frac{\beta^{R_s+R_{\bar{s}}} (1-\beta)^{X_s+X_{\bar{s}}-R_s-R_{\bar{s}}}}{1-(1-\beta)^{X_s+X_{\bar{s}}}} \times N! \frac{\alpha^{X_s+X_{\bar{s}}} (1-\alpha)^{N-X_s-X_{\bar{s}}}}{1-(1-\alpha)^N}. \end{aligned} \quad (6)$$

It should be noted that expression (6) is useful for setting up a probability model, but it is not actually the likelihood of the data unless the variables are completely observed. The appropriate likelihood of Bayesian inference for the actual information available is obtained by summing over the unknown quantities and not otherwise observed in the selected sample. The observed-data likelihood, conditional on λ , α and β , is given by:

$$p(\{i_1, \dots, i_m\}, X_s, R_s, \epsilon_s, \mathbf{C}_s, \mathbf{Y}_s) = \sum_{\mathbf{Y}_{\bar{s}}} \sum_{\mathbf{C}_{\bar{s}}} \sum_{\epsilon_{\bar{s}}} \sum_{R_{\bar{s}}} \sum_{X_{\bar{s}}} p(\{i_1, \dots, i_m\}, X, R, \epsilon, \mathbf{C}, \mathbf{Y}).$$

2.1 Prior distributions

In a Bayesian framework, the three unknowns α , β and $\boldsymbol{\lambda}$ are regarded as having been drawn from appropriate prior distributions. Assume that these parameters are independent; then, the joint prior distribution of $(\alpha, \beta, \boldsymbol{\lambda})$ is the product of their marginal prior distributions, described here. The parameter α controls the expected number of non-empty grid cells, and β controls the conditional expected number of non-empty networks. Figure 1 presents an illustration with certain artificial populations generated by model (2) and certain values fixed for α and β . We also arbitrarily fixed $\lambda_j = 10$, for all $j = 1, \dots, R$; thus, approximately 10 observations are expected in each unit.

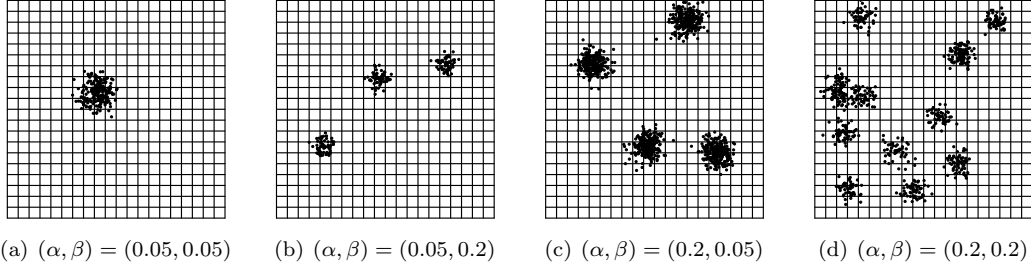


Figure 1: *Artificial populations generated by the proposed model with some fixed values for α and β , and $\lambda_j = 10$, for all $j = 1, \dots, R$, in a regular grid with $N = 400$ units.*

Because our approach aims to survey sparse populations, when analyzing Figure 1, it is reasonable to assume that both the α and β parameters should typically be small. To be uninformative with respect to these parameters, we should choose flat prior distributions. However, we can assign prior distributions that incorporate our knowledge of a rare and clustered population. In particular, we can consider that $\alpha \sim \text{Beta}(a_\alpha, b_\alpha)$ and $\beta \sim \text{Beta}(a_\beta, b_\beta)$ and choose values for the Beta distribution's parameters such that α and β are within an interval centered on a small value with high probability. The symbol $W \sim \text{Beta}(a, b)$ generically denotes that W is beta distributed and parameterized with mean $a/(a+b)$ and variance $ab(a+b+1)^{-1}(a+b)^{-2}$.

To ensure identifiability, it is necessary to adopt a unique labeling. For the proposed model in (2), unique labeling can be achieved by imposing a restriction on $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_R)'$. However, it should be noted that $\boldsymbol{\lambda}$, although totally unknown, has components associated with the sample where better estimates are expected. Thus, let us define $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_s, \boldsymbol{\lambda}_{\bar{s}})'$, such that $\boldsymbol{\lambda}_s$ refers to the networks observed in the sample and

$\lambda_{\bar{s}}$ to the unobserved networks. Note that it is necessary to impose a restriction on λ to ensure the identifiability of the model. Nevertheless, this restriction is only necessary for the elements of λ associated with the unknown networks, i.e., $\lambda_{\bar{s}}$.

Let us assume the following for λ :

$$\lambda \mid \theta \sim p(\cdot \mid \theta, R), \text{ such that } \lambda_j < \lambda_{j+1}, \text{ for all } j \in [R_s + 1, R_s + R_{\bar{s}}],$$

where $p(\cdot \mid \theta, R)$ represents the prior distribution of λ , which depends on the number of networks in the population, R , and on the vector of hyperparameters θ .

We use two different prior distributions for λ . First, we assume that the λ_j 's are conditionally independent given θ each with prior density $p(\lambda_j \mid \theta)$. Then, the joint prior density for λ is given by the following:

$$p(\lambda \mid \theta, R) = R_{\bar{s}}! p(\lambda_1 \mid \theta) \dots p(\lambda_R \mid \theta), \text{ such that } \lambda_j < \lambda_{j+1}, \text{ for all } j \in [R_s + 1, R_{\bar{s}}].$$

In particular, we consider $\lambda_j \sim \text{Gamma}(d, \nu)$, $\theta = (d, \nu)$ and introduce an additional hierarchical level by allowing ν to follow a $\text{Gamma}(e, f)$. The symbol $W \sim \text{Gamma}(a, b)$ generically denotes that W is gamma distributed and parameterized with mean a/b and variance a/b^2 .

One standard way of setting a Gamma as a weakly informative prior is to choose small values for its two parameters. However, such a distribution has a peak in the neighborhood of zero, which might encourage the inclusion of components with very small Poisson parameters, which would be difficult to estimate in general. Therefore, we used a weakly informative prior based on [17]-i.e., $\text{Gamma}(d, \nu)$ with d greater than one-to avoid the exponential shape without overly reducing the coefficient of variation's (CV) distribution. The parameter ν is set such that the prior mean d/ν is equal to the midrange of the observed data. However, in our case, we also consider ν to be unknown, so we choose e and f in the prior of ν such that the approximation to the mean of λ_j , $d/(e/f)$, is equal to the midrange of the observed data and the variance e/f^2 is relatively small.

The other prior considered for λ is the one introduced by [12] for normal mixtures as an explicit way to place an informative prior on the distance between two consecutive

λ_j 's. Here, the hyperparameter θ is τ , a positive constant, and the prior model is given by the following:

$$p(\boldsymbol{\lambda} \mid \tau, R) = p(\lambda_R \mid \lambda_{R-1}, \tau) p(\lambda_{R-1} \mid \lambda_{R-2}, \tau) \dots p(\lambda_1),$$

where $p(\lambda_j \mid \lambda_{j-1}, \tau)$ is $N_{(\lambda_{j-1}, \infty)}(\lambda_{j-1}, \tau)$, i.e., a Normal centered at λ_{j-1} with variance τ^2 , truncated to be greater than λ_{j-1} and $p(\lambda_1) \propto 1$. This ordering ensures the identifiability of the model.

[17] illustrate the difficulty of eliciting τ and its clear influence on the posterior distribution of the mixture parameters, as well as on the posterior distribution of the number of components. For example, if τ is very small compared to the anticipated distance between two consecutive λ_j 's, there will be a tendency to fit intermediate components between the true ones and hence to find a posterior distribution favoring higher values of R . This strategy gives a low prior probability that any two neighboring components are more than τ standard deviations apart. Based on a simulation study, [12] recommend choosing $\tau = 5$ because this choice leads to reasonable density estimates.

3 Inference

The posterior distributions of the parametric vector $\boldsymbol{\Theta} = (X_{\bar{s}}, R_{\bar{s}}, \boldsymbol{\epsilon}_{\bar{s}}, \mathbf{C}_{\bar{s}}, \mathbf{Y}_{\bar{s}}, \alpha, \beta, \boldsymbol{\lambda}, \nu)$ of model (2) cannot be obtained in closed form. Therefore, it is necessary to use some numerical approximation methods. One alternative, which is often used and is feasible to implement, is to generate samples from the marginal distributions of the parameters based on the Markov Chain Monte Carlo (MCMC) algorithm. Nevertheless, this method, as originally formulated, requires the posterior distribution to have a density with respect to some fixed measure. Thus, it cannot be used alone in this case, where the size of the parametric space is also a parameter. We use an approach based on reversible jump MCMC (RJMCMC), which was first proposed in [5] and applied in mixture models with unknown numbers of components by [11]. The method basically consists of jumps between the parameter subspaces corresponding to different numbers of components in the mixture.

For the proposed model (2), we used the steps specified below:

- (1) update the parameters α , β , $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$;
- (2) update the unobserved variables $X_{\bar{s}}$ and $\mathbf{Y}_{\bar{s}}$;
- (3) update the allocation $\boldsymbol{\epsilon}_{\bar{s}}$ so that $\mathbf{C}_{\bar{s}}$ is updated; and
- (4) combine two networks into one, or split one into two.

Steps (1)-(3) are performed using the Gibbs sampler or a Metropolis-Hastings sampler, and they do not change the dimensions of $\boldsymbol{\Theta}$. It should be noted that, because the proposed model (2) is defined only for the non-empty units, it is not possible to update the allocation, resulting in networks without any observations. Consequently, this step needs to be restricted so that each network must have at least one observation.

Step (4) involves changing $R_{\bar{s}}$ by 1 and making the necessary corresponding changes to $(\boldsymbol{\lambda}, \mathbf{C}, \boldsymbol{\epsilon})$. We made a random choice between splitting and combining, with probabilities depending on $R_{\bar{s}}$. Let $\lambda'_j = \lambda_j / \{1 - \exp(-\lambda_j)\}$ be the mean of the truncated Poisson distribution. The combination proposal begins by choosing a pair of components (j_1, j_2) at random, such that $\lambda'_{j_1} < \lambda'_{j_2}$. These two components are merged, forming a new component j^* . Now, we have to reallocate all the observations with $\epsilon_i = j_1$ or $\epsilon_i = j_2$ and create values for $(w_{j^*}, \lambda'_{j^*})$. They are chosen such that

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2}, \\ w_{j^*} \lambda'_{j^*} &= w_{j_1} \lambda'_{j_1} + w_{j_2} \lambda'_{j_2}, \end{aligned}$$

and we must impose $\lambda'_{j-1} < \lambda'_{j_1} < \lambda'_{j_2} < \lambda'_{j+1}$. A component j^* is chosen at random and split into j_1 and j_2 . However, there are two degrees of freedom for achieving this step, so we need to generate a two-dimensional random vector $\mathbf{u} = (u_1, u_2)$ to specify the new parameters. [17] present some ways of proposing a split that enforces the positivity constraint on Poisson parameters. In this work, we used the one referenced as “SM2” in their paper. In particular, the proposed model (2) is applicable to non-empty networks; thus, the split proposal also requires that both networks have at least one observation. Therefore, networks with only one observation cannot be chosen to be split. The acceptance probability for the split and combination steps can be viewed in Appendix A.

Although the expression above can be written in terms of λ'_j , the likelihood is expressed in terms of λ_j . Therefore, after step (4), we need to obtain λ_j from λ'_j by solving the equation $\lambda'_j = \lambda_j / \{1 - \exp(-\lambda_j)\}$. Furthermore, although the target function is invertible, it involves a polynomial with an exponential function, for which, in general, it is impossible to obtain an exact analytical solution. When the value of λ_j is sufficiently large, we can approximate λ_j by λ'_j (see Figure 2). However, for cases in which this approximation is not good, we need to use a numerical approximation, such as the Taylor approximation.

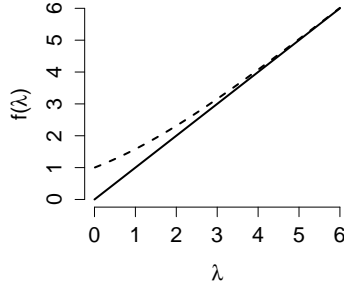


Figure 2: Comparison of the first-order moments of the Poisson distribution (—) and of the Poisson distribution truncated at zero (---).

3.1 Convergence diagnostics

To assess the performance of the proposed model and to check the convergence of the RJMCMC estimation, we generated a clustered population in an area with $N = 400$ units, fixing $\alpha = 0.15$ and $\beta = 0.10$. The values of the components of $\boldsymbol{\lambda}$ were generated from a Gamma distribution centered in 8.5 with a coefficient of variation (CV) equal to 95%, resulting in a Gamma distribution with parameters $d = 1.1$ and $\nu = 0.13$. Then, we selected a 5% sample using the adaptive design. We considered the prior distributions described in Section 2.1. For α and β , we chose $a_\alpha = 3$, $a_\beta = 15$, $b_\alpha = 1$ and $b_\beta = 9$, which parallel the prior distributions considered by [10]. These values are suitable when the only knowledge that can be obtained about the underlying population is that it is sparse and clustered. For $\boldsymbol{\lambda}$, we considered only the Gamma independent prior used in the generation of the artificial data. The population generated yields $R = 8$ networks

and, the networks observed were $s = \{2, 4, 7\}$, labeled such that the components of λ are in increasing order.

For the RJMCMC simulations, we generated 100,000 samples from the posterior distribution, discarded the first 10,000, and then thinned the chain by taking every 90th sample value. Figure 3 displays the histogram with the posterior densities of α , β , ν , λ and T for the generated population. The posterior densities of λ 's components are conditional on the posterior samples, whose estimated value of R is equal to eight. The solid and the dashed lines represent the true value and the 95% highest posterior density (HPD) interval, respectively. It should be noted that most of the parameters are well estimated, with their true value within the 95% HPD interval.

It should be noted that some λ_j 's associated with unobserved networks have bimodal posterior distributions and lower precision. This behavior is something expected in the posterior densities of mixture model parameters obtained by RJMCMC and is generally associated with the labeling at each sweep-see [11]. For instance, let us consider the case of two normal distributions, unambiguously labeled. The posterior distribution of the two means could overlap, but the extent of the overlap depends on its separation and the sample size. When the means are well separated, labels of the realizations from the posterior by ordering their means generally coincide with the population ones. As the separation reduces, "label switching" may occur. This problem can be minimized by choosing to order other parameters of the mixture components, for example, the variance. In our case, this bimodality does not appear in all the simulations, only on ones generated by the λ_j 's that are not well separated. Nevertheless, the bimodality influences neither the convergence of the other parameters nor the most important quantity: the total T .

The λ_j 's associated with the sampled networks present better estimates than the λ_j 's associated with the non-sampled networks. This result is expected because we have specific information for the sampled networks.

Two other diagnostics were used to show that the convergence was achieved: the Geweke and the Raftery-Lewis. The first was proposed by [4] and is based on a test for equality of the means of the first and last part of the Markov chain. If the samples are drawn from the stationary distribution of the chain, the two means are equal, and

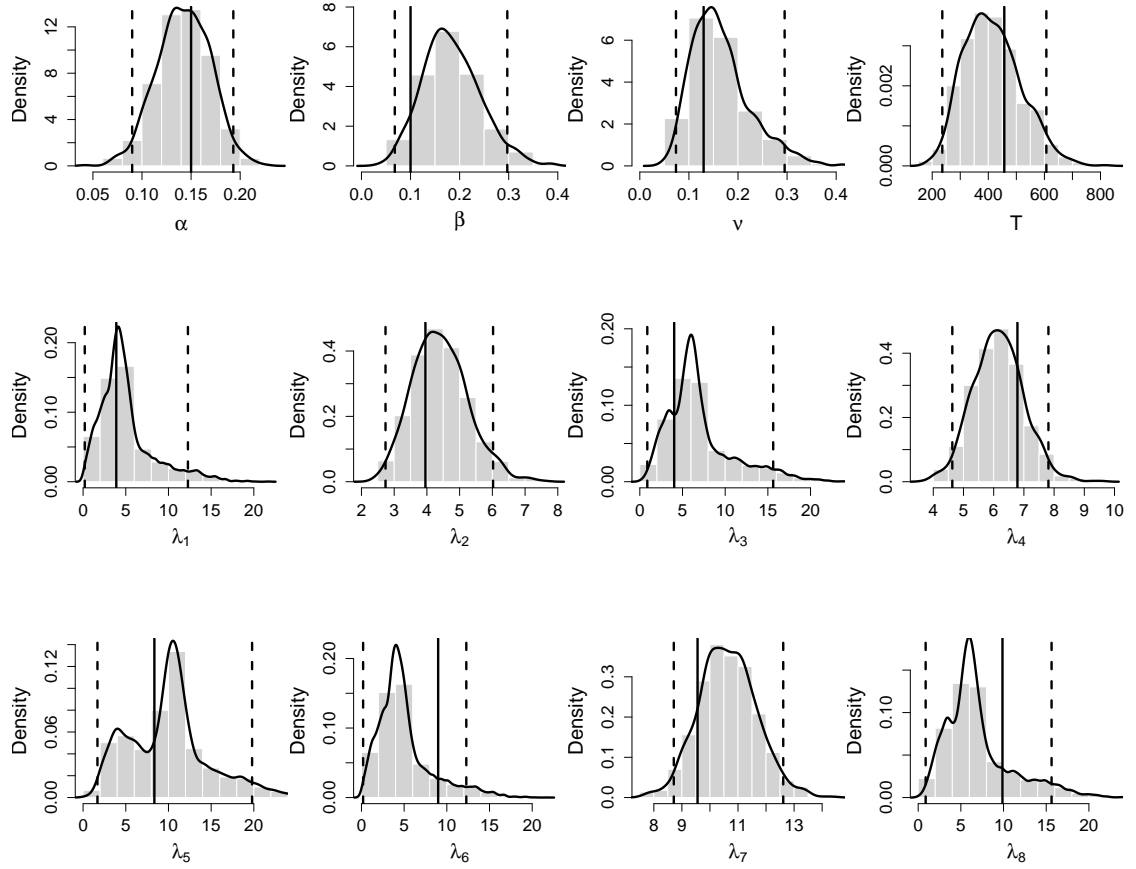


Figure 3: *Posterior densities for certain model parameters and the population total T for an artificial population. The vertical solid line is the true value fixed in the simulation, and the dashed line is the 95% HPD interval.*

Geweke's statistic has an asymptotically standard normal distribution. The second was proposed in [9] and calculates the number of iterations required to estimate a quantile with a desired accuracy and with a certain probability. The minimum length is the required sample size for a chain with no correlation between consecutive samples. An estimate dependence factor of the extent to which autocorrelation inflates the required sample size is also provided. Values for the factor that are larger than 5 indicate strong autocorrelation, which may be due to a poor choice of starting value, high posterior correlations or stickiness of the MCMC algorithm. Table 1 presents the value of Geweke's statistic and the value of the dependence factor. The results for both criteria indicate that the MCMC chains have converged.

Table 1: *Geweke and Raftery-Lewis convergence diagnostics for all of the parameters estimated for the artificial population.*

	α	β	ν	T	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
Geweke	0.7	-0.4	-1.6	0.4	1.4	-1.3	1.4	-0.4	1.5	1.5	1.2	1.5
R-L	1.3	1.1	1.1	1.8	0.9	1.0	1.0	1.0	0.9	1.0	1.1	1.1

4 Simulation study

To examine the performance of the Bayesian estimator and the influence of the different prior models on the Poisson parameters, we sampled several simulated clustered populations and obtained samples from the posterior distributions of the model parameters and population parameters. The population estimates were then compared with the true values to evaluate the model's performance.

4.1 Simulation scenarios

We generated 500 populations for each scenario that we considered. Twelve scenarios were created by varying the values of N , R and X , as well as varying the λ components. The values of parameters (α, β) were fixed such that their combinations expressed different degrees of rare and clustered populations. For the first simulation study, we considered only the independent prior for λ ; thus, we generated the values of the components of λ as a Gamma distribution with $d = 1.1$ and $\nu = 0.13$. These values of d and ν ensure that the generated populations provide heterogeneous networks. Finally, an adaptive cluster sample was selected from each population, with the first stage as a 5% simple random sample without replacement.

Table 2 shows summary statistics with some frequentist measures of the posterior distributions of the model parameters after reaching convergence for each of the twelve evaluated scenarios. It reports the relative mean square error (RMSE), the relative absolute error (RAE), the empirical nominal coverage of the 95% HPD intervals measured in percentages and the respective widths averaged over the 500 simulations. In particular, to facilitate future comparisons, the widths presented for the total T and for λ_s and $\lambda_{\bar{s}}$ are expressed in ratio form relative to their true values. The results for λ_j 's are separately summarized for λ_s and $\lambda_{\bar{s}}$.

Table 2: *Summary measurements for the point and interval estimates of the model and population parameters over 500 simulations for different values of α , β and N .*

	N = 200											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.21	0.38	0.53	0.56	0.03	0.29	0.22	0.29	0.29	0.39	0.03	0.28
RAE	0.35	0.17	0.25	0.60	0.12	0.46	0.36	0.16	0.35	0.47	0.13	0.45
Cov.	95.0	91.1	96.7	89.5	91.7	87.8	93.8	93.7	98.1	89.7	90.3	87.7
Wid.	1.60	0.20	0.31	0.28	0.58	1.23	1.60	0.19	0.31	0.28	0.57	1.26
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.09	0.20	0.50	0.22	0.02	0.31	0.06	0.10	0.19	0.32	0.02	0.27
RAE	0.24	0.31	0.45	0.40	0.11	0.46	0.21	0.27	0.21	0.47	0.10	0.41
Cov.	94.6	90.9	97.1	90.2	93.6	89.1	97.3	97.0	98.5	90.5	94.1	89.8
Wid.	1.22	0.19	0.21	0.22	0.50	1.33	1.24	0.20	0.23	0.21	0.56	1.51
	N = 400											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.06	0.15	0.42	0.14	0.02	0.29	0.05	0.08	0.15	0.10	0.02	0.31
RAE	0.21	0.32	0.35	0.28	0.10	0.43	0.20	0.23	0.29	0.21	0.12	0.43
Cov.	96.7	91.1	96.0	90.8	94.2	91.0	96.8	95.1	98.1	90.5	94.3	91.8
Wid.	1.04	0.09	0.20	0.19	0.47	1.38	1.05	0.10	0.21	0.18	0.55	1.64
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.04	0.06	0.35	0.04	0.02	0.30	0.05	0.03	0.15	0.03	0.02	0.36
RAE	0.18	0.18	0.39	0.18	0.09	0.42	0.20	0.15	0.21	0.15	0.10	0.43
Cov.	93.4	91.2	96.9	96.7	94.2	93.9	92.4	97.0	98.7	96.5	93.5	95.6
Wid.	0.79	0.11	0.15	0.14	0.45	1.43	0.77	0.11	0.16	0.13	0.51	1.77
	N = 600											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.04	0.05	0.25	0.10	0.02	0.32	0.05	0.03	0.11	0.09	0.02	0.35
RAE	0.17	0.17	0.28	0.12	0.09	0.42	0.20	0.14	0.26	0.11	0.11	0.42
Cov.	96.3	91.8	98.1	98.0	93.5	93.1	92.8	97.5	98.3	97.0	93.8	96.1
Wid.	0.79	0.08	0.22	0.20	0.46	1.40	0.78	0.08	0.23	0.19	0.52	1.70
	$(\alpha, \beta) = (0.15, 0.10)$						$(\alpha, \beta) = (0.15, 0.15)$					
RMSE	0.05	0.04	0.21	0.06	0.01	0.37	0.09	0.08	0.06	0.05	0.02	0.35
RAE	0.19	0.17	0.30	0.09	0.09	0.44	0.29	0.24	0.18	0.09	0.10	0.43
Cov.	90.4	91.1	98.7	98.9	95.3	96.0	90.0	90.5	98.8	98.4	95.5	96.8
Wid.	0.78	0.08	0.17	0.18	0.43	1.49	0.53	0.08	0.20	0.17	0.53	1.79

In general, the parameters are well estimated. The coverage of the 95% HPD intervals is close to the nominal level. The RMSE and RAE are small for all the parameters, except for β in certain specific cases. However, there is no significant impact on the prediction of the total T , which is our main interest. As expected, the results for λ_j obtained with the samples containing the network j show smaller errors and are more precise than the results that consider the samples in which the network j was not observed. As the value of N increases, the RMSEs and RAEs of most of the parameters decrease. This phenomenon may occur because the number of non-empty networks increases with N , improving the estimates of α and β and consequently of the other

parameters. However, for the same reason, for a fixed value of N , the errors decrease as the values of α and β increase.

It is not possible to present the frequentist properties for each λ_j because the value of R was not fixed over the simulations. Figure 4 presents the relative errors (REs) of λ_s and $\lambda_{\bar{s}}$ for all the networks and all the simulations, for different values of α and β and for $N = 400$. Note that, in all cases, the RE is approximately zero and is smaller for λ_s , as expected. Note also that $\lambda_{\bar{s}}$ is slightly underestimated.

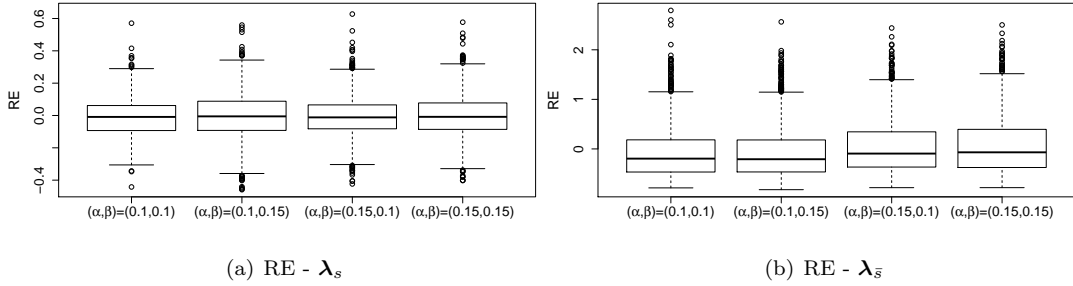


Figure 4: *Relative errors for λ_s and $\lambda_{\bar{s}}$ over 500 simulations, for $N = 400$ and different values of α and β .*

The 500 populations were previously generated by fixing the parameters of λ_j 's Gamma distribution at $d = 1.1$ and $\nu = 0.13$, yielding a mean of 8.5 and a CV of 95%. The aim here is to evaluate the performance of the proposed model with respect to the level of homogeneity. We consider two extra values of CVs: 25% and 50%, with the means fixed at 8.5 for both. Then, we calculate the two sets of values of d and ν . When the CV is fixed at 50%, we obtain $d = 4$ and $\nu = 0.47$; when the CV equals 25%, the result is $d = 16$ and $\nu = 1.89$.

Figure 5 displays the densities of λ_j for each fixed value of the CV. Note that, as the CV decreases, the prior distribution for λ_j becomes more concentrated and symmetrical around the mean of the distribution; consequently, the networks will become more homogeneous with respect to the total in their units.

We generated two other sets of populations, fixing the CVs of the λ_j distributions to 50% and 25%, respectively. The population size was set at $N = 400$, and a 5% adaptive sample was taken from it.

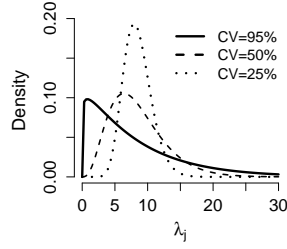


Figure 5: *Prior distributions to λ_j used in the simulations, varying the value of the CV of the distribution.*

Table 3 presents summary measurements of the estimators over the 500 populations generated for the two values considered for the CV. It should be noted that, even for the more homogeneous cases, the proposed model (2) has a good performance, resulting in parameter estimates with small errors and 95% HPD intervals with coverage probability near the fixed nominal level.

Table 3: *Summary measurements for the point and interval estimation of the model parameters over 500 simulations, varying the level of homogeneity in λ , expressed as the coefficient of variation of its distribution, for $N = 400$.*

	$CV = 50\%$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.13	0.15	0.52	0.16	0.02	0.04	0.06	0.09	0.18	0.10	0.02	0.03
RAE	0.26	0.32	0.27	0.30	0.10	0.15	0.18	0.24	0.36	0.23	0.11	0.15
Cov.	95.3	87.2	97.0	95.3	94.7	97.0	96.7	95.0	98.2	95.0	94.5	97.6
Wid.	1.38	0.11	0.26	0.91	0.51	1.27	1.24	0.11	0.27	0.82	0.55	1.31
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.03	0.04	0.40	0.08	0.02	0.03	0.03	0.03	0.10	0.06	0.02	0.03
RAE	0.15	0.15	0.50	0.21	0.10	0.12	0.16	0.14	0.26	0.18	0.10	0.13
Cov.	96.5	94.7	97.3	97.8	95.6	98.0	95.8	97.3	98.0	97.5	95.8	97.9
Wid.	0.95	0.11	0.23	0.75	0.48	1.28	0.92	0.11	0.24	0.70	0.53	1.36
	$CV = 25\%$											
	$(\alpha, \beta) = (0.10, 0.10)$						$(\alpha, \beta) = (0.10, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.09	0.30	0.50	0.36	0.03	0.08	0.05	0.18	0.12	0.34	0.03	0.08
RAE	0.23	0.48	0.37	0.47	0.13	0.24	0.19	0.37	0.29	0.44	0.14	0.26
Cov.	89.7	86.8	98.0	75.0	85.7	82.2	94.7	90.1	98.2	74.9	85.7	81.0
Wid.	0.96	0.12	0.25	3.01	0.47	0.70	0.91	0.12	0.27	2.83	0.51	0.75
	$(\alpha, \beta) = (0.15, 0.1)$						$(\alpha, \beta) = (0.15, 0.15)$					
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$	T	α	β	ν	λ_s	$\lambda_{\bar{s}}$
RMSE	0.03	0.08	0.41	0.25	0.02	0.03	0.04	0.05	0.07	0.19	0.02	0.04
RAE	0.14	0.22	0.49	0.34	0.10	0.15	0.17	0.15	0.21	0.24	0.11	0.17
Cov.	96.6	91.7	97.5	80.8	94.6	94.4	91.9	92.5	98.3	83.2	93.3	94.8
Wid.	0.70	0.12	0.22	2.48	0.46	0.74	0.70	0.12	0.23	2.25	0.50	0.79

In particular, the relative errors of T do not vary much with the values of the CV, except when $(\alpha, \beta) = (0.10, 0.10)$, for which, on average, smaller numbers of non-empty networks in the generated populations are found. In addition, the relative errors for $\lambda_{\bar{s}}$ are smaller than the errors obtained when CV is fixed in 95%, though the errors for ν become larger. Furthermore, as the CV decreases, the empirical coverage of nominal 95% HPD intervals is underestimated, mainly with respect to ν and λ .

4.2 Prior sensitivity analysis

In this section, we compare the performance of the two prior distributions considered for λ . To obtain simulation results for each component λ_j of λ using a different method from the previous section, the values of R were fixed. The population size was set at $N = 400$, and $(\alpha, \beta) = (0.15, 0.10)$. These settings were chosen to provide rare and clustered populations as much as possible. Then, we conducted a large number of simulations until we reached 500 populations with $R = 5$; another 500 populations were generated with $R = 6$, followed by another 500 populations with $R = 7$. We consider only these values of R because the others have much lower probabilities of being generated in this simulation scenario with $(\alpha, \beta) = (0.15, 0.10)$. Furthermore, because we were specifying two different priors for λ , we fixed the λ 's components at $(4.5, 4.8, 8.0, 11.3, 13.8)$ for $R = 5$, at $(3.9, 6.4, 6.9, 7.1, 10.5, 14.8)$ for $R = 6$ and at $(4.8, 7.4, 9.5, 10.1, 11.4, 11.7, 14.5)$ for $R = 7$. These values were generated from a uniform distribution defined in the interval $(3, 15)$.

All results shown hereafter correspond to 100,000 RJMCMC sweeps, after 10,000 burn-ins; the chain was then thinned by taking every 10^{th} sample value. We used the same prior distribution for α and β described in the previous section. For λ , we considered the Gamma prior distribution used in the previous simulation study and the dependent prior $\lambda_j \mid \lambda_{j-1} \sim N_{(\lambda_{j-1}, \infty)}(\lambda_{j-1}, \tau)$ with $\tau \in \{1, 5, 10, 20\}$.

Figure 6 shows the 95% HPD interval obtained for R for each λ prior assumed when we fit the model for one of the 500 populations generated. The parameter R is much more sensitive to the value of τ assigned for the dependent prior. In addition, the R posterior distribution is fairly vague when $\tau = 1$. However, as τ increases, this behavior is attenuated. The independent prior and the dependent one with $\tau = 20$

yield approximately the same 95% HPD interval for R . This behavior was observed for almost all of the 500 simulation samples. Thus, from now on, we do not consider the dependent λ prior with $\tau = 1$.

Figure 7 presents the RMSE for each λ_j display for samples where the network j is observed (a) and when it is not (b) for the four λ priors employed. Figure 7 shows that the independent prior provides a smaller RMSE than the dependent one for most cases, noticeably for the smaller λ_j 's. These results do not depend heavily on the values of τ . As expected, the RMSE values of the λ_j whose network j is not sampled are greater than the RMSE values of λ_j , for $j \in s$.

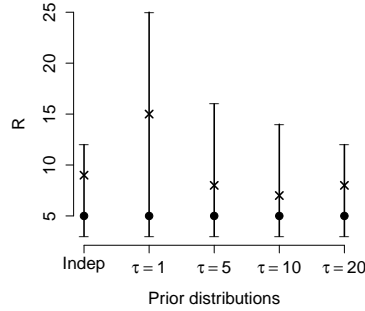


Figure 6: The 95% HPD interval of R for different prior distributions of λ . The crosses represent the median of the distribution, the circle represents the true value of R , and the line represents the 95% interval.

Because total population prediction is the main aim in this context, we also evaluate the impact of those prior distributions on the posterior distribution of T . Figure 8 displays the RMSE of T , the nominal coverage of the 95% HPD interval and its respective width for each considered value of R . We can observe from Figure 8 that the RMSEs obtained using the independent λ prior are always smaller than the ones obtained using the dependent λ priors. However, the 95% HPD intervals based on the dependent λ priors have higher coverage than the nominal level and higher width than when using the independent λ prior. Note that, for a fixed value of R , the results provided by the dependent priors are very similar for all values of τ .

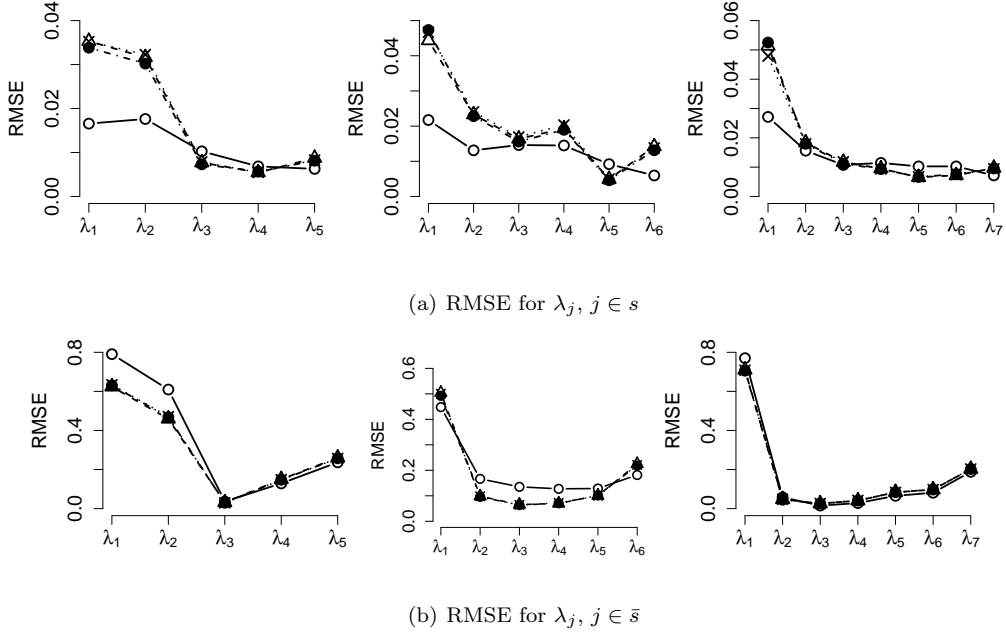


Figure 7: *RMSE of each λ_j assuming different priors for λ . The results with the independent prior distribution and the dependent ones with $\tau = 5$, $\tau = 10$ and $\tau = 20$ are respectively represented by the empty circles and the line, the triangles and the dashed line, the cross and the dotted line, and the full circle and the dot-dashed line.*

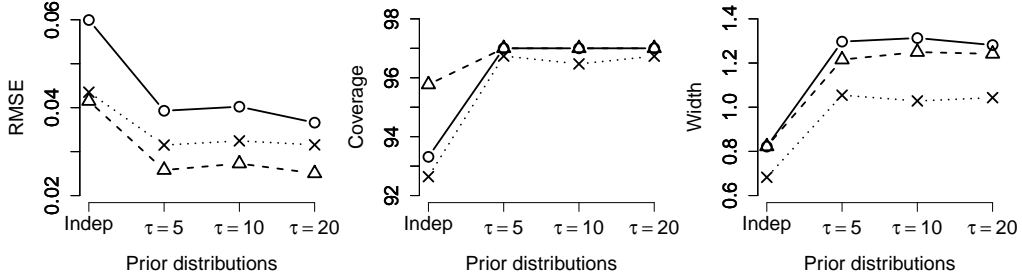


Figure 8: *RMSE, coverage and widths of the 95% HPD interval for the population total T for each prior distribution assumed for λ and for each R fixed. The results for $R = 5$, $R = 6$ and $R = 7$ are represented by the empty circles and the line, the triangles and the dashed line, and the cross and the dotted line, respectively.*

5 Comparison with the network model

The mixture model (2) has been presented as an alternative to that of [10]. The mixture model (2) is principally useful when we cannot assume homogeneity between networks with respect to the number of observations inside them and when the expected number

of observations inside any network is not proportional to its respective area size. The key idea of this paper is to improve on the population estimates obtained by [10] through the use of a model that takes into account heterogeneity between networks. This is accomplished by modeling at the unit level rather than at the network level.

To assess the effectiveness of our methodology, we compared the results of our approach to the results obtained in [10]. The first comparison consists of a design-based experiment with a real population, and the second study is a model-based experiment. To fit both models, we assigned the same prior distributions used in Subsection 4.1. To conduct the MCMC and RJMCMC simulations, we generated two chains of length 100,000 each, discarded the first 10,000 and then thinned the chain by taking every 90th sample value to obtain 1,000 independent samples.

5.1 A design-based experiment

We evaluated the proposed model (2) by performing a design-based experiment in which adaptive samples were drawn from a real, fixed population. Design-based studies are used in the context of survey sampling inference to evaluate the performance of model-based estimators under repeated samples taken from a real, fixed population where a characteristic of interest is known for all its units. This real population can be a Census or a large sample that is supposed for evaluation purposes to be the population. The main aim of this design-based experiment is to analyze the frequentist properties of the total estimators using both approaches.

The population used here for design-based evaluation is the same described in [14] and consists of counts of a waterfowl species, called the blue-winged teal, in a $5,000 \text{ km}^2$ area of central Florida in 1992. Figure 9 shows the counts of blue-winged teals in a grid with $N = 200$ units. It should be noted that these counts are sparse and clustered, justifying the use of adaptive sampling.

The study consists of selecting 500 adaptive samples with initial sizes within 10% from the population. From now on, we will refer to the model of [10] as the ‘network model’. Note that the assumptions of their model are not wholly suitable for the blue-winged teal data. Nevertheless, our proposed model assumes heterogeneity among units, which seems more reasonable when we analyze Figure 9. Furthermore, note that there

Table 4. We calculated the ratio of the variances of both Bayes estimators and referred to it as efficiency (ef) in Table 4.

Table 4 shows that the network model presents larger errors than our proposed model (2). The network model produces credible intervals that, despite their larger width, have a lower nominal coverage than desired. Furthermore, our proposed model is more efficient when applied to these data.

Table 4: *Summary measurements for the point and interval estimates of the total population, obtained by fitting the proposed and the network models.*

	RMSE	RAE	Coverage	Width	ef(\hat{T})
Mixture model	0.01	0.05	96.7	0.25	0.87
Network model	0.03	0.13	85.6	0.35	

Figure 10 shows the boxplots with the REs for the population’s total posterior means and true values based on the 500 samples when fitting both models. Here again, we see that the REs obtained for our proposed model are lower, although both overestimate the true values. This result is not unexpected, as there is a network with a substantially different number of observations from the others.

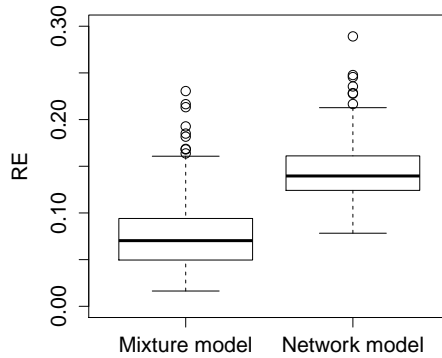


Figure 10: *Boxplots with the RE for T for the 500 samples, obtained by the fits of the proposed and network models.*

5.2 A model-based experiment

The purpose of this simulation study is to compare the performance of the network and mixture models when the populations are generated according to the mixture model. We considered two scenarios. For the first scenario, we used the same populations of

500 generated in the simulation study presented in Section 4.1 and fitted the network model to evaluate its performance. In particular, we considered the case where $(\alpha, \beta) = (0.15, 0.10)$. For the second scenario, we generated the components of $\boldsymbol{\lambda}$ according to a Gamma distribution with CV=25%. Thus, it was expected that the network model performance would improve because the homogeneity degree of $\boldsymbol{\lambda}$'s components was higher than in the first scenario (CV=50%).

Table 5 displays some frequentist properties of the estimators obtained by fitting the network model. To facilitate the comparison, the results when fitting the mixture model with the same populations are presented in Table 5 in parentheses. Regarding the estimation of T , both models have equivalent performance when CV=25%. However, as the degree of homogeneity decreases, the mixture model performs considerably better than the network model. However, the network model exhibits better performance than the mixture model with respect to the parameter β for both scenarios.

Table 5: *Summary measurements for the point and interval estimates of the network model parameters over 500 simulations where $\boldsymbol{\lambda}$ were generated from a Gamma distribution with CV=25% and 50%, for $N = 400$ and $(\alpha, \beta) = (0.15, 0.10)$.*

	CV=25%			CV=50%		
	T	α	β	T	α	β
RMSE	0.03 (0.03)	0.05 (0.08)	0.18 (0.41)	0.05 (0.03)	0.04 (0.04)	0.10 (0.40)
RAE	0.17 (0.14)	0.16 (0.22)	0.32 (0.49)	0.21 (0.15)	0.19 (0.15)	0.37 (0.50)
Cov.	96.8 (96.6)	97.1 (91.7)	95.6 (97.5)	95.6 (96.5)	98.1 (94.7)	97.4 (97.3)
Wid.	0.86 (0.70)	0.16 (0.12)	0.19 (0.22)	0.85 (0.95)	0.16 (0.11)	0.18 (0.23)

Finally, we present the boxplot of the relative error of T for both models in Figure 11. The conclusion is analogous to the other measurements. In particular, the estimator provided by the network model seems to underestimate T for both scenarios.

Therefore, from those results, it should be concluded that as the level of homogeneity between networks increases, the performances of the evaluated models become similar. The main difference is the number of parameters to estimate and the computational effort, which is more significant when fitting the mixture model.

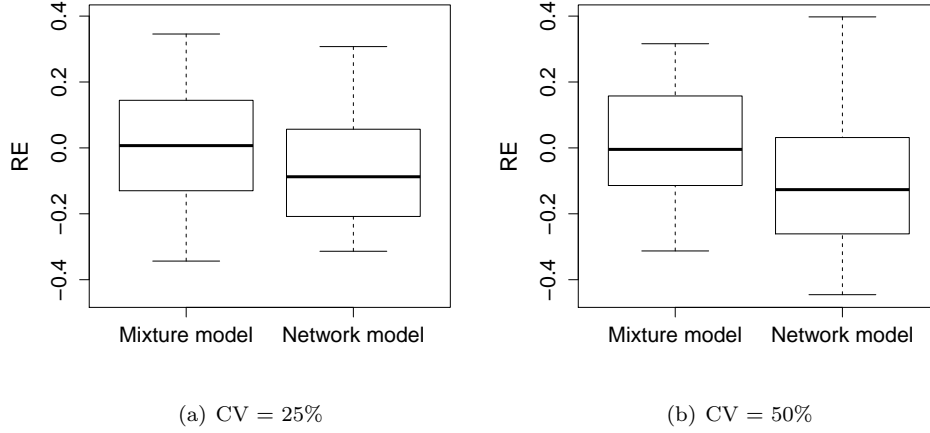


Figure 11: *Boxplots with the RE for T for the 500 samples, obtained by the fits of the proposed and the network models, for a Gamma distribution for λ with CV=25% and CV=50%.*

6 Conclusions and suggestions for future work

We have considered the problem of estimating the total numbers of individuals in a rare and clustered population. Our approach is to model the observed counts in grid cells, selected by adaptive cluster sampling, and then to use model-based analysis to estimate the total population. The proposed model is an alternative to that of [10] because it models the grid cells instead of the networks and supposes heterogeneity between units that belong to different networks. Nevertheless, it requires considerable computational effort and should therefore be used only if the data support it. However, simulation studies show that as homogeneity between networks decreases, it might be worth using the mixture model as an alternative to the network model.

More general assumptions can be considered and modeled within this framework. For example, in the same network, units near the centroid should have higher frequency than units that are far from the centroid. It is possible to consider this assumption in the proposed model.

It should be noted that the parameters of the response variable associated with the unobserved components present some estimation difficulties. Therefore, the prior distribution should be carefully elicited. Thus, the main findings of this work encourage an extension of the model-based analysis to other adaptive sampling plans, which uncover more information about the population. One example is adaptive cluster

double sampling, proposed by [2], which allows the sampler to control the number of measurements of the variable of interest and to use auxiliary information.

Acknowledgements

This work is part of the Ph.D. thesis of Kelly C. M Gonçalves under the supervision of Fernando Moura, in the Graduate Program of UFRJ. Kelly has a scholarship from Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES). Fernando Moura receives financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil, BPPesq). The authors would like to thank the editor, an associate editor and two referees for their very thoughtful and constructive comments.

A Acceptance probability for the split or combination moves

For the split step, to obtain the acceptance probability it is necessary to simulate (u_1, u_2) from distributions with densities g_1 and g_2 , respectively. The probability of acceptance, supposing an independent prior distribution for λ , is $\min(1, A)$, where:

$$\begin{aligned}
A = & \frac{\exp\{-(c_{j_1}\lambda_{j_1} + c_{j_2}\lambda_{j_2})\} \lambda_{j_1}^{\sum_{i:\epsilon_i=j_1} y_i} \lambda_{j_2}^{\sum_{i:\epsilon_i=j_2} y_i} (1 - \exp(-\lambda_{j_1}))^{-c_{j_1}} (1 - \exp(-\lambda_{j_2}))^{-c_{j_2}}}{\exp\{-c_{j^*}\lambda_{j^*}\} \lambda_{j^*}^{\sum_{i:\epsilon_i=j^*} y_i} (1 - \exp(-\lambda_{j^*}))^{-c_{j^*}}} \\
& \times \frac{p(\{i_{j_1}, i_{j_2}\})}{p(\{i_{j^*}\})} \times \frac{p(R_{\bar{s}} + 1)}{p(R_{\bar{s}})} \times \frac{(c_{j^*} - 1)!}{(c_{j_1} - 1)!(c_{j_2} - 1)!} (R_s + R_{\bar{s}})^{-(c_{j_1} + c_{j_2} - c_{j^*})} \times \frac{c_{j_1}^{c_{j_1}} c_{j_2}^{c_{j_2}}}{c_{j^*}^{c_{j^*}}} \times (R_{\bar{s}} + 1) \\
& \times \frac{\nu^d}{\Gamma(d)} \left(\frac{\lambda_{j_1} \lambda_{j_2}}{\lambda_{j^*}} \right)^{d-1} \exp\{-\nu(\lambda_{j_1} + \lambda_{j_2} - \lambda_{j^*})\} \\
& \times \frac{p_{R_{\bar{s}}+1|R_{\bar{s}}}}{p_{R_{\bar{s}}+1|R_{\bar{s}}} P_{alloc} q(u_1) q(u_2)} \times |J|,
\end{aligned}$$

where $p_{R_{\bar{s}}+1|R_{\bar{s}}}$ is the probability of choosing the split step, P_{alloc} is the probability that this particular allocation is made, and $|J|$ is the Jacobean of the transformation $(w_{j^*}, \lambda'_{j^*})$ to $(w_{j_1}, w_{j_2}, \lambda'_{j_1}, \lambda'_{j_2})$. For the corresponding combination step, the acceptance probability is $\min(1, A^{-1})$, and simple adaptations must be made because the proposal reduces the number of nonsampled networks by 1.

B Assessment of MCMC and RJMCMC with real data

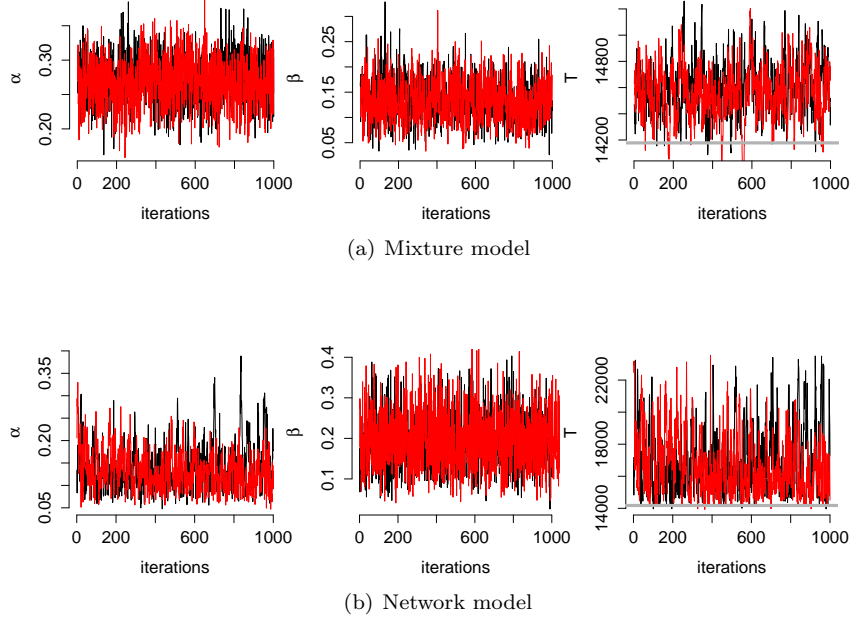


Figure 12: Trace plot with the posterior densities of α , β and T obtained by the fits of the proposed and the network models. The gray line represents the true value of T .

Table 6: Geweke and Raftery-Lewis convergence diagnostics for some of the parameters estimated for the real population for both models.

Param	Geweke		Raftery-Lewis	
	Mixture	Network	Mixture	Network
α	-0.13	-0.10	1.02	1.21
β	0.72	-0.67	1.15	2.56
T	-1.38	-0.30	3.22	1.33

References

- [1] Connors, M. and Schwager, S. (2002) The use of adaptive cluster sampling for hydroacoustic surveys. ICES Journal of Marine Science: Journal du Conseil, **59**, 1314-1325.

- [2] Felix-Medina, M. H. and Thompson, S. K. (2004) Adaptive cluster double sampling. *Biometrika*, **91**, 877-891.
- [3] Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) Bayesian data analysis. Chapman & Hall.
- [4] Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments." In J. Bernardo, A. D., J. Berger and Smith, A. (eds.), *Bayesian Statistics*. Oxford University Press, New York.
- [5] Green, P. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711-732.
- [6] McDonald, L. L. (2004) Sampling rare populations. In Sampling rare or elusive species: concepts, designs, and techniques for estimating population parameters (ed. W. Thompson), chap. 4, 11-42. Island Press Washington, DC, USA.
- [7] Pfeiffermann, D., Moura, F. A. D. S. and Silva, P. L. D. N. (2006) Multi-level modelling under informative sampling. *Biometrika*, **93**, 943-959.
- [8] Philippi, T. (2005). "Adaptive cluster sampling for estimation of abundances within local populations of low-abundance plants." *Ecology*, 86(5): 1091-1100.
- [9] Raftery, A. E. and Lewis, S. M. (1992). "One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo." *Statistical Science*, 7(4): 493-497.
- [10] Rapley, V. and Welsh, A. (2008) Model-based inferences from adaptive cluster sampling. *Bayesian Analysis*, **3**, 717-736.
- [11] Richardson, S. and Green, P. (1997) On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, **59**, 731-792.
- [12] Roeder, K. and Wasserman, L. (1997) Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894-902.

- [13] Roesch, F. (1993) Adaptive cluster sampling for forest inventories. *Forest Science*, **39**, 655-669.
- [14] Smith, D., Conroy, M. and Brakhage, D. (1995) Efficiency of adaptive cluster sampling for estimating density of wintering waterfowl. *Biometrics*, **51**, 777-788.
- [15] Thompson, S. K. (1990) Adaptive cluster sampling. *Journal of the American Statistical Association*, **85**, 1050-1059.
- [16] Thompson, S. K. and Seber, G. A. F. (1996) *Adaptive sampling*. Wiley New York.
- [17] Viallefont, V., Richardson, S. and Green, P. J. (2002) Bayesian analysis of poisson mixtures. *Journal of Nonparametric Statistics*, **14**, 181-202.